Convergence Doesn't Show Lexically-Specific Phonetic Detail

Chelsea Sanker
Yale University

1 Introduction

Can phonetic convergence be lexically specific, or does convergence occur only at a phonological level? Some studies find more convergence in lower frequency words, which is taken as evidence for word-specific phonetic detail (Goldinger, 1998; Babel, 2010; Nielsen, 2011; Dias & Rosenblum, 2016). However, this is only indirect evidence for word-specific convergence and other studies have failed to replicate the effect (e.g. Pardo et al. 2013, 2017). Word-specific convergence would provide strong evidence for word-specific representations that include phonetic details, if it can be clearly demonstrated to exist. I present two studies that probe whether convergence can differ by word and whether previously observed frequency effects might have an alternative explanation.

1.1 *Frequency effects* Several studies have found a correlation between lexical frequency and convergence, with more convergence in lower frequency words (Goldinger, 1998; Babel, 2010; Nielsen, 2011; Dias & Rosenblum, 2016). This effect is usually explained within Exemplar Theory, as laid out by Goldinger (1998). For lower frequency words, the exemplars from the task are a large proportion of the overall cloud of weighted exemplars, producing strong convergence. For higher frequency words, there are more pre-existing recent exemplars, so the exemplars from the task have a smaller impact in shifting that robust representation.

However, different behaviors based on lexical frequency are an indirect test of word-specific convergence. It is possible that there is a correlation between lexical frequency and apparent convergence driven by something other than word-specific phonetic detail. Some studies have sought to replicate the correlation between lexical frequency and convergence and not found evidence for it (e.g. Pardo et al. 2013, 2017), which casts further doubt on the effect. A greater increase in similarity between speakers in lower frequency words than in higher frequency words does not necessarily reflect frequency-conditioned differences in convergence.

It is possible that apparent frequency effects in convergence could be produced as a result of how repetition effects interact with frequency. The first mention of words is prone to hyperarticulation, while utterances of the same words become more natural in subsequent productions (Bard et al., 2000; Fowler & Housum, 1987). Repetitions of these words are likely to continue to increase in naturalness over several repetitions. For this reason, some convergence studies establish speakers' baselines by measuring their second production of a word (e.g. Nielsen 2011; Pardo et al. 2013). However, many studies take baselines using a single production of each word.

The effect of first mention is likely to be stronger for lower frequency words, particularly when produced in frame sentences or in isolation, which provides no contextual predictability; lower frequency words take longer to read and elicit more errors than higher frequency words (Gerhand & Barry, 1998; Beattie & Butterworth, 1979). Wright (1979) demonstrates that duration decreases with repetition more for low frequency words than for high frequency words. Thus, speakers' later productions of low frequency words are likely to differ from their initial productions more than high frequency words do. The particular effects of repetition might also differ for high frequency words, because they are likely to exhibit more reduction in subsequent productions. High frequency words might exhibit more shifts in vowel formants and other characteristics strongly influenced by reduction, while low frequency words are more influenced by characteristics like duration and f0 that are strongly influenced by hesitations and other disfluencies.

Natural fluent productions will be more similar across speakers than hyperarticulated or disfluent productions, so increased naturalness could create the appearance of convergence when the shift is compared to another speaker as a reference value. Because low frequency words are likely to be more hyperarticulated

Proceedings of *AMP 2020*

and contain more disfluencies in their initial productions than higher frequency words are, increased naturalness will tend to be greater for lower frequency words, and thus lower frequency words will seem to exhibit more convergence.

Any initial production that is an outlier is likely to result in apparent convergence, because a more typical subsequent production will on average be more similar to most other speakers. Cohen Priva & Sanker (2019) demonstrate that the difference-in-difference method of measuring convergence is highly susceptible to this artifact of regression to the mean, because any change in difference from the model talker is treated as convergence or divergence, even though it is sometimes due to random variation, effects of the task, or other factors. If baselines are unreliable because they are consistently skewed by hyperarticulation and disfluencies, this can produce artifacts not only in difference-in-difference measurements but also in holistic AXB evaluations of similarity; two natural productions of a word are likely to sound more similar than one atypical production and a different atypical production or a natural production.

As examples, I will consider the methods of four shadowing studies that found lexical frequency to be a significant predictor of convergence. Goldinger (1998) and Dias & Rosenblum (2016) used AXB evaluations of similarity to measure convergence. Babel (2010) used difference-in-difference to measure convergence. Nielsen (2011) measured change in VOT, which should not be as susceptible to the issue of unreliable baselines. She also used each speaker's second production of a word as a baseline, which further reduces the risk of unreliable baselines. Stimuli had manipulated VOT, so convergence would be reflected in lengthened or shortened VOT in each task, respectively; there was only significant convergence to lengthened VOT, and only an effect of frequency in this condition. It is possible that the results reflect a different repetition effect than what is described above. The baseline VOT productions are likely to be longer than speakers' typical productions, which could reduce some of the naturally produced variation in VOT. Once listeners settle into natural productions of the now-familiar words, the typical VOT for higher frequency words will be shorter than the VOT of lower frequency words, producing a greater spread of VOT across words than was present in the baselines. While in normal productions this spread is the result of shortening in high frequency words, the overall VOT targets are lengthened due to the shadowing stimuli, so the ultimate VOT for high frequency words might look similar to the baseline VOTs, while the VOT for lower frequency words is longer.

1.2 *Consistency and variation across words* The frequency effects in previous convergence studies are interpreted as reflecting the accumulation of exemplars of each word, based on words having their own individual exemplar clouds that contain phonetic details specific to that word. If word-specific phonetic details exist, there should also be other evidence for such word-specific details.

Previous work observes convergence not only in the particular words that a participant is exposed to during the task but also in other words, which demonstrates that shared phonological representations are active in convergence. However, extension to new words when stimuli are consistent does not necessarily indicate that word-specific learning would be impossible when supported by distinctions in the stimuli. The characteristics investigated within a study are usually consistent across the stimuli, either naturally produced by the same speaker (e.g. Goldinger 1998), or manipulated in the same way, such as with lengthened VOT in all items (e.g. Nielsen 2011). It is possible that extension only occurs when stimuli are consistent, providing listeners with evidence that the pattern should be generalized because it is present across all items.

Generalizations at the phonological level may coexist with word-specific phonetic details, as is proposed in hybrid exemplar models (e.g. Pierrehumbert 2002). In such models, each exemplar of a word is both part of the cloud of exemplars for the particular word and also part of the cloud for each sound in the word, which also includes all other words that contain that sound. Thus, if listeners hear enough exemplars of a particular word with acoustic characteristics that consistently deviate from the typical realization of the component sounds, it should be possible for the target pronunciation of that word to become distinct from other words, even if they have the same phonological elements.

Homophones are a key part of the lexicon where word-specific phonetic details should be clear, if they exist. It has been demonstrated that homophone mates can exhibit significant differences in their acoustic details as they are produced in natural speech (e.g. Gahl 2008; Lohman 2018), which could indicate that they have distinct phonetic details in their representations. However, many of the differences in production can be attributed to factors such as position in the sentence (Conwell, 2017) and predictability in context (Jurafsky et al., 2002). The differences between homophone mates can be largely eliminated when they are produced in frame sentences or in isolation (Guion, 1995; Sanker, 2019), which suggests that the acoustic

differences are merely an effect of context, rather than being part of the representation. Perception results similarly suggest that listeners do not have distinct phonetic details in the representation of homophone mate pairs. Bond (1973) found at chance accuracy for identifications of homophone mates. While Sanker (2019) found accuracy slightly above chance for identification of homophone mates in some conditions, the effect was small. Moreover, slightly above chance accuracy might be explained by expectations based on frequency, without listeners needing word-specific phonetic representations.

Listeners are sensitive to the prototypicality of the acoustic details of the stimuli, at least at the phonological level. Andruski et al. (1994) demonstrate that words manipulated to have less prototypical VOT in a word-initial stop produce weaker priming than words with VOT typical of the phonological category. In a word identification task, Scarborough & Zellou (2013) demonstrate that response times are longer for hyperarticulated speech than naturally produced speech, suggesting that more prototypical forms are easier to perceive accurately. This effect might also be word-specific, or at least involve knowledge of frequency-based reduction patterns; Scarborough and Zellou found that it interacts with neighborhood density. The difference between conditions was smaller for low neighborhood density words than high neighborhood density words. Given the high correlation between neighborhood density and frequency, this result might indicate that hyperarticulated low frequency words are more similar to the prototypical forms of those words than hyperarticulated high frequency words, which are rarely produced this way in natural speech.

If listeners have word-specific phonetic representations, it should be possible to shifts the acoustic targets in different ways for different words. Such shifts have been observed with altered auditory feedback. Rochet-Capellan & Ostry (2011) demonstrate that exposing subjects to increased F1 in "bed" and decreased F1 for "head" resulted in word-specific compensatory shifts: decreased F1 in "bed" and increased F1 in "head." However, just because such a manipulation is possible does not mean that it has been elicited in convergence studies. One crucial way that Rochet-Capellan and Ostry's study differs from convergence studies is in the amount of exposure to each word. Listeners only heard three words during the task ("bed", "head", "ted"), each presented over 100 times. In contrast, most convergence studies have a much larger number of words and exponentially fewer repetitions of each word; some of the most extensive exposure is used by Goldinger (1998), who has a condition with 12 repetitions of each item. The nature of the tasks might also produce differences. In convergence, exposure is meant to shift the acoustic details in the representation; in altered feedback studies, exposure is meant to shift productions while the representation remains unchanged.

2 Study 1

Study 1 examines whether effects of repetition can produce the appearance of frequency-conditioned convergence, by testing if participants' second production of each word is more similar to other participants' productions than their first production was.

2.1 *Methods* 24 female native speakers of American English participated in the study. Having only female participants was done in order to avoid the issue of normalization, which could produce substantially different predictions for cross-gender pairs than same-gender pairs.

The participants read a set of 120 monosyllabic English words twice in randomized order. They were not exposed to any other speaker during this task. Words were selected so that their log transformed lexical frequencies were approximately normally distributed.

F1, F2, vowel duration, and f0 mean were measured for each production of each word. These measurements were then analyzed to test whether participants were more similar to each other in their second productions than in their first productions. In a shadowing experiment or other study that involved exposure to other speakers, this would be a test of convergence. However, the participants did not hear any other speaker during the task, so this is a test of potential artificial convergence effects that actually result from repetition.

There were 24 iterations of each analysis, one using each participant's productions as the reference values, as if that participant had been a model talker heard by the others. In each model, the other 23 participants' pronunciations of each word were compared to the reference speaker's pronunciation. There were two versions of how the reference values were defined: one using the reference speaker's first productions of the words as the reference values, and one using the reference speaker's second productions of the words as the reference values. Both of these are paralleled by methods that are often used to create stimuli for convergence studies; sometimes the stimuli are made from a speaker saying each word a single time, and

	Estimate	*SE*	*t-value*	*p-value*
(Intercept)	0.039	1.3	0.031	0.98
log lex frequency	-1.1	0.38	-3.0	0.0031

Table 1: Summary of one of the mixed effects regression models for difference-in-difference in f0, using one participant's first productions as the reference levels.

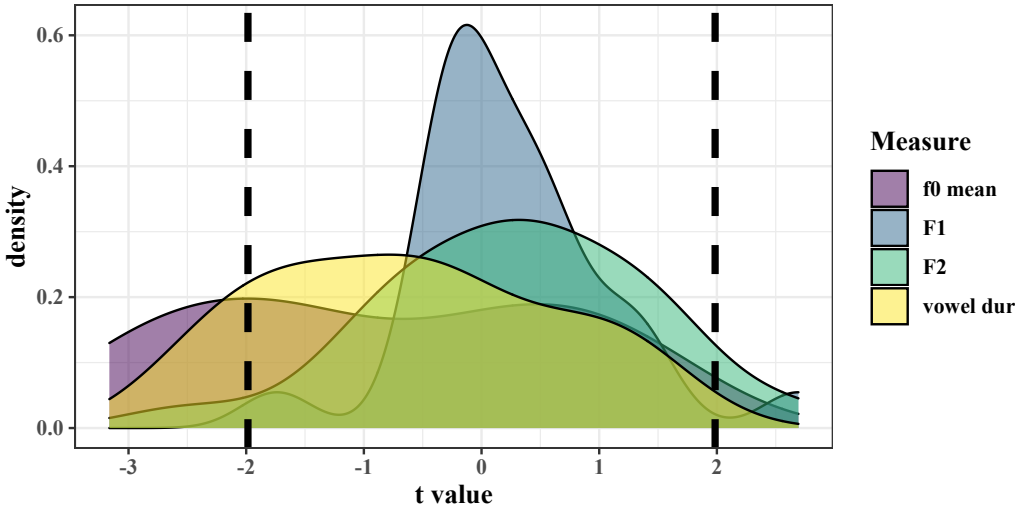


Figure 1: *t-values* for lexical frequency as a predictor of difference-in-difference from the reference speaker's first productions. Dashed lines indicate the significance threshold.

sometimes the words will be produced multiple times and the clearest or most natural items will be used as stimuli.

Following previous work on frequency effects, "convergence" was measured as change in distance from the reference value: $\mid ParticipantStart - Reference \mid$ - $\mid ParticipantFinal - Reference \mid$, also called difference-in-difference. Participants were measured as "converging" when their second production of a word was more similar to the reference speaker's production of that word than their first production was. These analyses were run for each of the 4 acoustic measurements, producing 192 total models (4 acoustic measurements * 24 reference speakers * 2 reference productions).

Statistics come from mixed effects regression models predicting convergence, calculated with the lme4 package in R (Bates et al., 2015), and *p-values* were calculated by the lmerTest package (Kuznetsova et al., 2015).

2.2 *Results* Table 1 provides an example of one of the 192 regression models used to analyze apparent "convergence." All models used log lexical frequency as the one fixed effect and included random intercepts for participant and word. Lexical frequency was centered, to improve interpretability of the intercept. The main element of interest in each model is the *t-value* for lexical frequency. Given that there was no possibility for convergence in this task, there also should not be any effect of lexical frequency in predicting convergence.

Although participants were not exposed to another speaker, repetition produced the appearance of frequency-conditioned convergence when participants were compared to each other. Figure 1 illustrates the *t-values* for lexical frequency as a predictor of change in distance from the reference speaker's first productions, across all models of this type. Because there cannot be any convergence, lexical frequency should not be a predictor of convergence, so these *t-values* should be centered at 0.

Given a significance threshold of 0.05, about 5 of the 96 models would be expected to exceed this threshold by chance, with no tendency to be positive or negative. However, the results instead indicate clear

patterns. Lexical frequency was a significant negative predictor, predicting more convergence with lower frequency words, in 12/96 models and a significant positive predictor in 2/96 models. There are also notably different distributions for each acoustic measure. Both F1 and F2 have *t-values* centered close to 0, while f0 and vowel duration have *t-values* that are distributed more broadly and centered more in negative ranges.

Figure 2 illustrates the *t-values* for lexical frequency as a predictor of change in distance from the reference speaker's second productions, across all models of this type. The results are similar to the the results when using the reference speaker's first productions: Lexical frequency was a significant negative predictor in 8/96 models and a significant positive predictor in 1/96, and again the negative estimates disproportionately come from models of f0 and vowel duration.

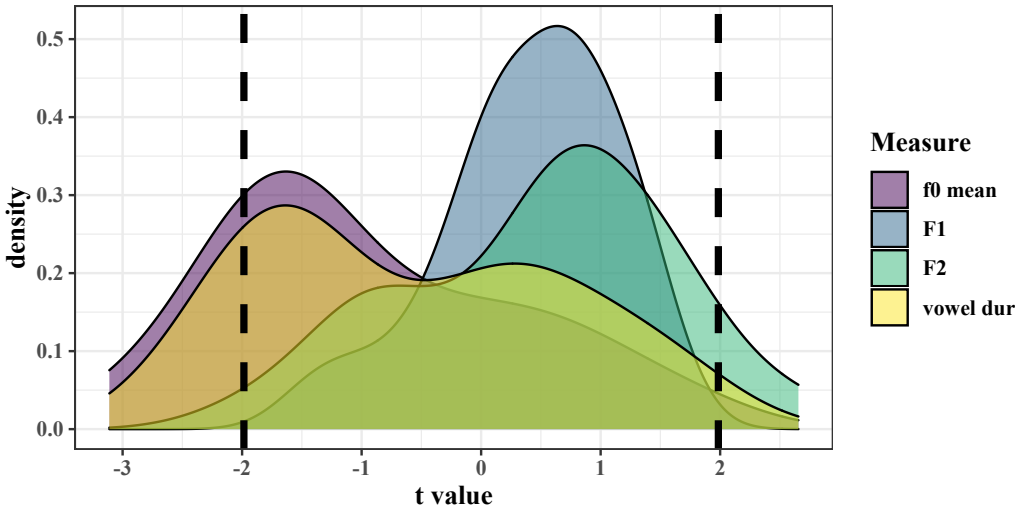


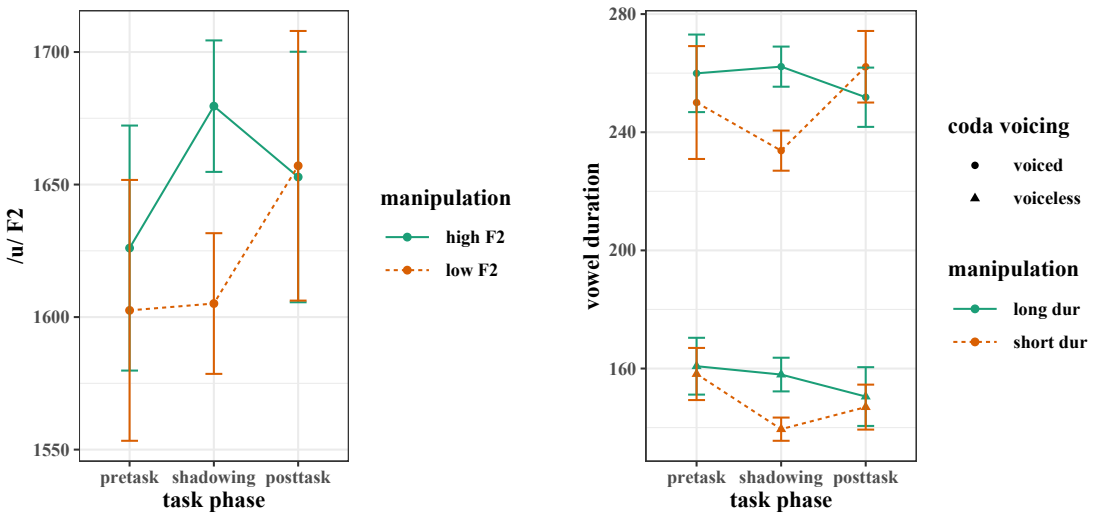
Figure 2: *t-values* for lexical frequency as a predictor of difference-in-difference from the reference speaker's second productions. Dashed lines indicate the significance threshold.

Given that repetitions seem to produce the appearance of frequency-conditioned convergence, it is also worth commenting on the overall appearance of convergence in these models. When using the reference speaker's first productions as the reference value, 2/96 models found significant convergence and 9/96 models found significant divergence. When using the reference speaker's second productions as the reference value, 7/96 models found significant convergence, and 6/96 models found significant divergence. These results suggest that repetition is altering these acoustic characteristics. The tendency towards finding divergence when comparing participants' productions to the first production of the reference speaker is probably driven by reduction; most speakers' second productions are more reduced than the first, making them more distinct from the reference speaker's first production. The mixed results when comparing participants to the second production of the reference speaker suggests that the different effects of repetition between higher frequency and lower frequency words can produce the overall appearance either of convergence or divergence, depending on which characteristic is being measured. The results would also be sensitive to the distribution of lexical frequencies present within the stimuli.

3 Study 2

The results of the preceding study cast some doubt on whether previously observed frequency effects in convergence studies necessarily reflect word-specific convergence. Study 2 aims to directly test word-specific convergence, by exposing listeners to words that have been manipulated in opposing directions.

3.1 *Methods* 24 female native speakers of American English participated in the study. The stimuli were made from recordings of a female native speaker of American English reading individual words in randomized order in a sound attenuated booth. As before, having only female participants and a female model talker was done in order to avoid the issue of normalization, which could produce substantially different predictions for cross-gender pairs than same-gender pairs.



(a) /u/ F2 **(b)** vowel duration

Figure 3: Means and 95% confidence intervals by task phase, characteristic measured, and manipulation

First, participants read a set of monosyllabic words twice in randomized order; this included 36 target words and 84 filler words. The second production was used as the baseline for each word.

In the shadowing task, participants repeated after the acoustically manipulated target words in randomized order. Each stimulus was presented three times. Two characteristics were investigated: F2 in /u/ and vowel duration. For each characteristic, each listener heard an equal number of words with each manipulation; the manipulation was always the same for the three repetitions of the same lexical item, e.g. *boot, brute, hoot, moose, shoes, zoo* with raised F2, and *boost, cooed, choose, do, fruit, hoop* with lowered F2. Items in each condition were balanced to have similar phonological environments. The characteristics being investigated were selected to include (1) a spectral measure that is inherent in English phonological contrasts but has variable realization in American English, and (2) a temporal measure that is not necessarily inherent in English phonological contrasts but has variable realizations based on several phonological and non-phonological factors.

After the shadowing task, participants again read the full set of words. Participants' vowel duration and F2 were measured for each word in each phase of the experiment. There were two stages of statistical analysis. First, the results were modelled with mixed effects regression models using the lme4 package in R (Bates et al., 2015), with *p-values* calculated by the lmerTest package (Kuznetsova et al., 2015). Second, the results were modelled with Bayesian regression models, using the brms package (Bürkner, 2017).

3.2 *Results* Figures 3a-3b illustrate the results for both measurements in each phase of the task. The pre-task productions were used to establish the speakers' baselines for each measurement, and confirm that there was no initial difference between the words in each condition. There is a clear separation in the acoustic characteristics based on the manipulation condition during immediate repetition in the shadowing phase, but no evidence for preservation of that effect in the post-task productions.

Table 2 presents the summary of a mixed effects regression model for F2 in /u/. The fixed effects were manipulation (high F2, low F2), phase (shadowing, pre-task, post-task), and the interaction between them. There were random intercepts for participant and for word.

The model confirms the significant effect of manipulation on F2 within the shadowing phase. Within immediate repetitions, F2 of /u/ was 74.5 Hz lower in the Low F2 condition than in the High F2 condition, i.e. participants produced a lower F2 when they were repeating after a word in which F2 was manipulated to have a low F2 than when repeating after a word manipulated to have a high F2.

There was a significant interaction between manipulation and task phase; the effect of manipulation on

	Estimate	*SE*	*t-value*	*p-value*
(Intercept)	1679.6	57.4	29.2	< 0.001
Manipulation LowF2	-74.5	11.5	-6.5	< 0.001
Phase Pretask	-77.8	16.9	-4.6	< 0.001
Phase Posttask	-51.0	16.9	-3.0	0.0026
Manip LowF2 : Phase Pretask	57.0	23.8	2.4	0.017
Manip LowF2 : Phase Posttask	84.74	23.8	3.6	< 0.001

Table 2: Summary of a mixed effects regression models for F2 in /u/. *Reference Levels: Manipulation = High F2, Phase = Shadowing*

	Estimate	*Est. Error*	*lower 95% CI*	*upper 95% CI*
Intercept	1646.8	67.3	1512.5	1779.4
Manipulation LowF2	12.2	21.4	-29.7	53.9

Table 3: Summary of a Bayesian mixed effects regression models for F2 in /u/ in post-task productions. *Reference Levels: Manipulation = High F2*

F2 was eliminated in the post-task productions. Because the model did not provide evidence for the two manipulation conditions producing a difference in post-task productions, the post-task data was analyzed with a Bayesian model, to examine the probability that there is no effect.

Table 3 presents the summary of a Bayesian mixed effects regression model for F2 in /u/ in the post-task phase. The fixed effect was manipulation (high F2, low F2). There were random intercepts for participant and for word. The intercept was given a broad normally distributed prior, with a mean of 1600 Hz and a standard deviation of 400, based on previous work on dialectally varied /u/ fronting in American English (e.g. Labov et al. 2008:Ch. 10). In order to avoid a possible bias of the prior in producing a model with no effect of manipulation, the model prior had a mean of -100 Hz and a standard deviation of 400.

The Bayesian model suggests that it is likely that there was no effect of the F2 manipulation on post-task productions. The credible interval is close to 0 and includes values above and below 0. Note also that the estimate is slightly positive for words with the Low F2 manipulation as compared to words with the High F2 manipulation, which is in the opposite direction of the effect that would be predicted by convergence.

Table 4 presents the summary of a mixed effects regression model for vowel duration. The fixed effects were manipulation (short duration, long duration), phase (shadowing, pre-task, post-task), coda voicing (voiced, voiceless), vowel (/æ, ei, i, u/), and the interaction between manipulation and phase. There were random intercepts for participant and for word.

The model confirms the significant effect of manipulation on vowel duration within the shadowing phase. Within immediate repetitions, vowels were 23.1 ms shorter in the Short Duration condition, i.e. participants produced shorter vowels when repeating after a word manipulated to have short duration than when repeating after a word manipulated to have a long duration.

There was a significant interaction between manipulation and task phase; the effect of manipulation on vowel duration was eliminated in the post-task productions. Because the model did not provide evidence for the two manipulation conditions producing a difference in post-task productions, the post-task data was analyzed with a Bayesian model, to examine the probability that there is no effect. This model also included an interaction between manipulation and coda voicing, to test whether the manipulation might have produced effects that differed based on coda voicing.

Table 5 presents the summary of a Bayesian mixed effects regression model for vowel duration in the post-task phase. The fixed effects were manipulation (short duration, long duration), coda voicing (voiced, voiceless), vowel (/æ, ei, i, u/), and the interaction between manipulation and coda voicing. There were random intercepts for participant and for word. The intercept was given a broad normally distributed prior, with a mean of 250 Hz and a standard deviation of 200. The prior for voiceless codas relative to voiced codas had a mean of -100 and a standard deviation of 50. The prior for each of the vowels relative to /æ/ had a mean

	Estimate	*SE*	*t-value*	*p-value*
(Intercept)	281.7	11.4	24.7	< 0.001
Manipulation ShortDur	-23.1	8.2	-2.8	0.01
Phase Pretask	-0.57	3.3	-0.17	0.86
Phase Posttask	-9.4	3.3	-2.9	0.0043
CodaVoicing Voiceless	-100.3	8.0	-12.5	< 0.001
Vowel ei	-17.8	10.7	-1.7	0.11
Vowel i	-28.2	10.7	-2.6	0.017
Vowel u	-37.1	10.7	-3.5	0.0028
Manip ShortDur : Phase Pretask	18.4	4.6	4.0	< 0.001
Manip ShortDur : Phase Posttask	23.8	4.6	5.1	< 0.001

Table 4: Summary of a mixed effects regression models for vowel duration. *Reference Levels: Manipulation = Long Duration, Phase = Shadowing, CodaVoicing = Voiced, Vowel =* /æ/

	Estimate	*Est. Error*	*lower 95% CI*	*upper 95% CI*
Intercept	273.4	13.6	247.1	299.2
Manipulation ShortDur	9.6	13.4	-16.5	36.2
Vowel ei	-19.1	11.8	-42.5	3.5
Vowel i	-33.7	12.1	-57.5	-9.7
Vowel u	-31.8	11.9	-55.4	-8.1
CodaVoicing Voiceless	-101.7	12.8	-127.3	-76.6
Manip ShortDur : CodaVoi Voiceless	-13.0	18.5	-49.8	23.5

Table 5: Summary of a Bayesian mixed effects regression models for vowel duration in post-task productions. *Reference Levels: Manipulation = Long Duration, CodaVoicing = Voiced, Vowel =* /æ/

of -20 and a standard deviation of 50. In order to avoid a possible bias of the prior in producing a model with no effect of manipulation, the model prior had a mean of -100 Hz and a standard deviation of 200.

The Bayesian model suggests that it is likely that there was no effect of the vowel duration manipulation on post-task productions. The credible interval is close to 0 and includes values above and below 0. Note that the estimate is slightly positive for words with voiced codas in the Short Duration condition relative to the Long Duration condition, which is in the opposite direction of the effect that would be predicted by convergence. The results also suggest that there is no interaction between manipulation condition and coda voicing; the effect of duration manipulation did not differ based on coda voicing.

4 Discussion

Study 1 finds a repetition-based increase in similarity to other speakers that is correlated with lexical frequency. Correlations between frequency and increased similarity have been used as evidence for word-specific convergence in previous work; however, the existence of the relationship in data from a task with no exposure to other speakers might indicate that the previously observed correlations are based on frequency-conditioned repetition effects in production rather than frequency-conditioned convergence. Because convergence studies involve multiple repetitions of each word, they are also likely to exhibit repetition effects, so a correlation between frequency and apparent convergence could show up as an artifact. Previous results finding correlations between lexical frequency and convergence thus do not provide clear evidence to support word-specific phonetic detail; phonetically detailed lexical representations are not necessary to explain effects of repetition.

The effect of repetition can probably be explained by fluency of first mentions of a word, particularly for words produced in isolation or in frame sentences, in which they are not predictable from the preceding context. Low frequency words have lower fluency than other words in initial productions, because they

have low predictability; these low frequency, low predictability words are more likely to exhibit disfluencies (Beattie & Butterworth, 1979; Gerhand & Barry, 1998) and are likely to be less reduced and more likely to be hyperarticulated (Hall et al., 2018). Subsequent productions are likely to be more fluent, and thus more similar to other speakers' productions. This explanation is supported by the behavior of the particular acoustic characteristics measured in this study. Vowel duration and mean f0 exhibited the clearest tendency towards this correlation between lexical frequency and increased similarity across speakers, and these are two primary characteristics that are strongly influenced by hesitations and other disfluencies (e.g. Shriberg 2001).

Words in this study were selected to have an approximately normal distribution of log frequencies, typical of the lexicon as a whole. However, convergence studies aimed at addressing frequency effects usually design the stimuli so that there is a set of high frequency words and a set of low frequency words (e.g. Goldinger 1998; Dias & Rosenblum 2016). Weighting the distribution towards high frequency and low frequency words is likely to increase the probability that the repetition effects will produce a significant correlation between frequency and apparent convergence, because there are more items at the extreme ends of the spectrum where the different impact of repetition effects will be most apparent. Thus, the results of Study 1 are probably underestimating how frequently this artifact will show up in convergence studies.

Study 2 tested whether word-specific convergence could be elicited by manipulating different words in different directions, using a shadowing paradigm similar to what has been used in previous convergence studies. This experiment provided no evidence for lexically-specific convergence. Acoustic details are reflected in immediate repetition during the shadowing phase, but this short-term effect could be explained by auditory traces. That is, listeners can mimic the specific acoustic details of words that they have just heard when they repeat those words, but this does not require the acoustic details to have entered their representations.

Neither F2 of /u/ nor vowel duration exhibited evidence for shifts in the representational target of the words. In a shadowing task, a shift in the representation is most clearly reflected in how listeners produce the words after the task, when they are drawing on their own representations rather than repeating after an acoustic stimulus. In this experiment, the was no evidence for an effect of the manipulation condition on post-task productions; the estimated posterior means for both characteristics were close to 0, and the credible intervals included values above and below 0.

Even if words have their own specific representations, they also have shared phonological representations, which are strongly established based on extensive input across words and across speakers (Pierrehumbert, 2002). In this study, the incoming stimuli disagree with each other in their phonetic details at the phonological level, so listeners cannot make a generalization about a shifted realization of the phonological category. A lack of lexically-specific details could produce the observed result because convergence is averaged across words with the same sound. Different words were manipulated in opposite directions, so convergence to the average characteristics of the phonological categories would not produce any change in the manipulated characteristics. It is also possible that listeners expect consistency across words and will not converge to an inconsistent speaker, based on perceiving that speaker as an unreliable source of input. Both explanations would reflect shared acoustic details at the phonological level, rather than word-specific representations.

Word-specific convergence in this task would require that listeners learn arbitrary new phonetic targets for each word individually. From previous work on homophones, it is not clear that people even learn non-arbitrary word-specific phonetic details. The acoustic differences between homophone mates that appear in natural speech are largely eliminated when homophones are produced in frame sentences or in isolation (Guion, 1995; Sanker, 2019), and listeners cannot reliably distinguish between homophone mates in perceptual identification tasks (Bond, 1973; Sanker, 2019). Apparent word-specific effects of prototypicality in setting perceptual expectations (e.g. Scarborough & Zellou 2013) might reflect knowledge about frequency-based reduction as a general pattern rather than requiring independent phonetic memories of each word. Expectations about frequency effects could also potentially result in homophone identifications that are slightly above chance, if the homophone mates differ substantially in frequency.

It is possible that the lack of word-specific convergence in this experiment is due to participants not having enough data to learn the pattern for each word. Establishing word-specific phonetic details may require substantial exposure to consistently distinct productions of these words. Work in altered auditory feedback (Rochet-Capellan & Ostry, 2011) has demonstrated that it is possible for participants to learn word-specific patterns of pronunciation, when the stimuli include a very large number of repetitions for a very small number of words. However, in Rochet-Capellan and Ostry's task, with only three words, it is also possible that listeners

made a generalization based on the phonological environment, rather than treating each pattern as word-specific; with a larger number of words, environmentally-conditioned generalizations would rapidly become untenable unless the stimuli were specifically designed to fall into natural groups. The internalization of shifted characteristics may also behave differently depending on how the shifts are elicited; in convergence, the actual target pronunciation changes in the representation, while in adaptation to altered feedback, the representation is unchanged but participants are influenced to alter how they try to achieve that target. While altered auditory feedback does show accumulated effects across trials of manipulation, the effects fade quickly, so Rochet-Capellan and Ostry's word-specific results may be more comparable to the convergent effects during shadowing, rather than the post-task results.

Even if word-specific convergence might be possible with extensive exposure to each word, it is likely that no study has elicited lexically specific convergence, because the amount of exposure in previous convergence studies is similar to the exposure in this study. If three repetitions of each word is not enough to produce even a weak shift for the words in each manipulation category, future convergence experiments examining word-specific effects need to be designed to account for the amount of exposure that is likely to be necessary.

5 Conclusion

Apparent frequency-conditioned convergence could be produced just based on repetition effects in production; a relationship between lexical frequency and increased similarity between speakers is apparent even in recordings of a word reading task, which involved no exposure to other speakers. This relationship could result from lower fluency for the first reading of lower frequency words, with subsequent more fluent productions being more similar to other speakers' productions. As frequency-conditioned convergence has been a major source of evidence for word-specific phonetic details, it may be valuable to also re-examine other evidence for such details.

When word-specific convergence is tested directly, with manipulations that differ by word, there is no evidence that speakers converge to these distinct forms. While this result does not demonstrate that word-specific acoustic details are impossible, it does suggest that these details, if they do exist, cannot be so easily established with limited exposure.

References

Andruski, Jean E., Sheila Blumstein & Martha Burton (1994). The effect of subphonetic differences on lexical access. *Cognition* 52, 163–187.

Babel, Molly (2010). Dialect divergence and convergence in New Zealand English. *Language in Society* 39, 437–456.

Bard, Ellen Gurman, Anne H. Anderson, Catherine Sotillo, Matthew Aylett, Gwyneth Soherty-Sneddon & Alison Newlands (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language* 42, 1–22.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1, 1–48.

Beattie, Geoffrey W & Brian L Butterworth (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22:3, 201–211.

Bond, Zinny S. (1973). The perception of sub-phonemic phonetic differences. *Language and Speech* 16:4, 351–355.

Bürkner, Paul-Christian (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80:1, 1–28.

Cohen Priva, Uriel & Chelsea Sanker (2019). Limitations of difference-in-difference for measuring convergence. *Laboratory Phonology* 10, Article 15.

Conwell, Erin (2017). Prosodic disambiguation of noun/verb homophones in child-directed speech. *Journal of Child Language* 44:3, 734–751.

Dias, James W. & Lawrence D. Rosenblum (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics* 78, 317–333.

Fowler, Carol A. & Jonathan Housum (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26, 489–504.

Gahl, Susanne (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* 84:3, 474–496.

Gerhand, Simon & Christopher Barry (1998). Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24:2, 267–283.

Goldinger, Stephen D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105:2, 251–279.

Guion, Susan G. (1995). Word frequency effects among homonyms. *Texas Linguistic Forum*, vol. 35, 103–116.

Hall, Kathleen Currie, Elizabeth Hume, T Florian Jaeger & Andrew Wedel (2018). The role of predictability in shaping phonological patterns. *Linguistics Vanguard* 4:s2, Article 20170027.

Jurafsky, Daniel, Alan Bell & Cynthia Girand (2002). The role of the lemma in form variation. Gussenhoven, Carlos & Natasha Warner (eds.), *Laboratory Phonology VII*, Mouton de Gruyter, Berlin, 3–34.

Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojesen Christensen (2015). *lmerTest: Tests in Linear Mixed Effects Models*. URL `https://CRAN.R-project.org/package=lmerTest`. R package version 2.0-29.

Labov, William, Sharon Ash & Charles Boberg (2008). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Lohman, Arne (2018). *Cut*(n) and *cut*(v) are not homophones: Lemma frequency affects the duration of noun-verb conversion pairs. *Journal of Linguistics* 54:4, 1–25.

Nielsen, Kuniko (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39, 132–142.

Pardo, Jennifer S., Kelly Jordan, Rolliene Mallari, Caitlin Scanlon & Eva Lewandowski (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language* 69, 183–195.

Pardo, Jennifer S., Adelya Urmanche, Sherilyn Wilman & Jaclyn Wiener (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics* 79, 637–659.

Pierrehumbert, Janet (2002). Word-specific phonetics. Gussenhoven, Carlos & Natasha Warner (eds.), *Laboratory Phonology VII*, Mouton de Gruyter, Berlin, 101–140.

Rochet-Capellan, Amélie & David J. Ostry (2011). Simultaneous acquisition of multiple auditory-motor transformations in speech. *Journal of Neuroscience* 31:7, 2657–2662.

Sanker, Chelsea (2019). Effects of lexical ambiguity, frequency, and acoustic details in auditory perception. *Attention, Perception, & Psychophysics* 81, 323–343.

Scarborough, Rebecca & Georgia Zellou (2013). Clarity in communication: "clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *Journal of the Acoustical Society of America* 134:5, 3793–3807.

Shriberg, Elizabeth (2001). To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31:1, 153–169.

Wright, Charles E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition* 7:6, 411–419.